Component of Statistics Canada Catalogue no. 12-002-X
Research Data Centres
ISSN: 1710-2197

Article

The Research Data Centres Information Information and Technical Bulletin

Fall 2012, vol. 5 no. 1

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca, email us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States)	
Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369
Local or international calls	
Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access this product

This product, Catalogue no. 12-002-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by "Key resource" > "Publications."

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2012

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement (http://www.statcan.gc.ca/reference/copyright-droit-auteur-eng.htm)

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
p	preliminary
r	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
E	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

About the Information and Technical Bulletin

The Research Data Centres Information and Technical Bulletin is a forum for current and prospective users of the centres to exchange practical information and techniques for analyzing datasets available at the centres. The bulletin is published twice per year, in the spring and fall. Additional special issues on timely topics may also be released on an occasional basis.

Aims:

The main aims of the bulletin are:

- to advance and disseminate knowledge surrounding Statistics Canada's data;
- to exchange ideas among the Research Data Centre (RDC) user community;
- to support new users of the RDC program; and
- to provide an additional means through which RDC users and subject matter experts and divisions within Statistics Canada can communicate.

Content:

The Information and Technical Bulletin is interested in receiving articles and notes that will add value to the quality of research produced at the Statistics Canada Research Data Centres and provide methodological support to RDC users.

Topics include, but are not limited to:

- data analysis and modeling;
- data management;
- best or ineffective statistical, computational, and scientific practices;
- data content;
- implications of questionnaire wording;
- comparisons of data sets;
- reviews on methodologies and their applications;
- problem-solving analytical techniques; and
- explanations of innovative tools, using surveys and relevant software available at the RDCs.

Those interested in submitting an article to the Information and Technical Bulletin are asked to refer to the Instructions for authors.

The editors and authors would like to thank the reviewers for their valuable comments.

Editor-in-Chief: Darren Lauzon
Managing Editor: James Chowhan
Associate editors: Jean-Michel Billette, Heather Hobson, and Georgia Roberts

Table of contents

Articles

Analyzing Census microdata in an RDC: What weight to use? 6
Georgia Roberts

Instructions for authors 13

Analyzing Census microdata in an RDC: What weight to use?

By Georgia Roberts[1]

Abstract

It is generally recommended that weighted estimation approaches be used when analyzing data from a long-form census microdata file. Since such data files are now available in the Research Data Centres (RDC's), there is a need to provide researchers with information about conducting weighted estimation with these files. The purpose of this paper is to provide some information – in particular, how the weight variables were derived for the census microdata files and what weight should be used for different units of analysis. For the 1996, 2001 and 2006 censuses the same weight variable is appropriate regardless of whether individuals, families or households are being studied. For the 1991 Census, the recommendations are more complex: a different weight variable is required for households than for individuals and families, and additional restrictions apply to obtain the correct weight value for families.

Introduction

"The census provides a statistical portrait of our country and its people. A vast majority of all countries regularly carry out a census to collect important information about the social and economic situation of the people living in its various regions. In Canada, the census is the only reliable source of detailed data for small groups (such as lone-parent families, ethnic groups, industrial and occupational categories, and immigrants) and for areas as small as a city neighbourhood or as large as the country itself."[2]

"The census includes every man, woman and child living in Canada on Census Day, as well as Canadians who are abroad, either on a military base, attached to a diplomatic mission, at sea or in port aboard Canadian-registered merchant vessels. Persons in Canada including those holding a temporary resident permit, study permit or work permit, and their dependents, are also part of the census."[2] However, only a limited amount of basic information is collected from everyone, while detailed data are requested only from a sample of the dwellings and people identified in the complete census enumeration. In 2001, for example, the census short-form questionnaire consisted of seven questions while each long-form questionnaire included the same seven questions from the short questionnaire and 52 additional multipart questions.

Weighted estimation techniques should be used when producing estimates from data from a long-form questionnaire file. As explained in more detail later, biased results can occur if suitable weights are not used. The purpose of this note is to provide background and justification for the weight variable(s) included on the census microdata files for 1991 to 2006, all of which, at the time of writing, are available in the Research Data Centres (RDC's). Section 2 briefly describes what records are in the census microdata files and the content of the records. Then, in Section 3, there is an explanation of the probabilities of inclusion of different units in the sample.

1. Georgia Roberts @ statcan.gc.ca , Data Analysis Resource Centre, Statistics Canada.

2. Extracted from http://www12.statcan.ca/census-recensement/2006/ref/about-apropos/faq-eng.cfm

This is followed by a review of the process of going from these probabilities to the specific weight variables that appear on the data files used by the researchers in the Research Data Centres. Section 4 gives the justification for fewer weight variables being on the RDC census files compared to the more extensive census databases at Statistics Canada Head Office. A final section provides concluding remarks.

What do the RDC census microdata files contain?

While the designation of who should provide long-form census information has changed slightly over time (and should be verified by the researcher through the documentation provided for each census), for the censuses from 1991 and 2006 it generally consisted of:

1. all individuals in a one-in-five sample of occupied private households in self-enumeration EAs (EAs were called collection units, or CUs, in 2006),[4]
2. all individuals[4] in non-institutional collective dwellings,[5]
3. all non-institutional individuals living in institutional collective dwellings,
4. all individuals in all occupied private households in canvasser area EAs – which are mainly northern and remote areas and Indian reserves,
5. all Canadians posted abroad, such as federal and provincial public servants and members of the Armed Forces, as well as Canadian citizens outside Canada who ask to be included in the census.

The content of the long-form questionnaires provided to these different groups of individuals were very similar. The main differences were in the omission of housing questions for some groups and an adaptation of examples to questions on the questionnaires distributed in northern areas and on Indian reserves.

Long-form microdata files containing person-level records have been created for the RDC's from the data from each of the 1991 to 2006 censuses, with files from other censuses possibly to follow. Each person-level record includes identifiers (such as household and family identifiers), geographic variables, and direct and derived variables from the questionnaire. Each record contains data from the questions common to both short-form and long-form questionnaires as well as data from the additional questions included just on the long-form questionnaires. Also on each person-level record are one or more survey weight variables to be used for estimation of quantities of interest to the researcher.

The population targeted by the whole long-form sample of each of the four censuses considered here can be described as all non-institutional "usual" residents of Canada (in or outside Canada), landed immigrants and non-permanent residents at the time of the census.

3. The one-in-five sampling fraction has been in place since 1951, except in 1971 and 1976, when it was one in three. The 1941 Census was the first census where detailed data were collected from a sample of households and, for that census, a one-in-ten fraction was sampled.

4. An "individual" is a Canadian citizen, a landed immigrant or a non-permanent resident.

5. The 2006 Census was slightly different from the other 3 censuses. In particular, people in shelters (which are non-institutional collective dwellings) received only short-form questionnaires in 2006. As well, "senior units" were introduced to the long-form sample in 2006, as described in the 2006 Census Technical Report: Sampling and Weighting (Statistics Canada, 2009).

Institutionalized Canadians cannot be studied using these long-form microdata files since institutionalized Canadians received only short-form questionnaires and records containing the basic information collected about them is not included in these files. (Institutional residents, formerly known as inmates of institutions, include "nonstaff" residents of prisons, jails, hospitals, nursing homes, etc.) A researcher may be interested in estimating descriptive quantities of the entire target population, or descriptive statistics about a subpopulation, or more complex statistics such as model coefficients.

How are survey weight variables developed for a census microdata file?

For the development of a survey weight variable for a long-form census microdata file, first to be determined is the probability of inclusion in the sample of each particular unit. Because, as described in the previous section, a straightforward plan was used to choose who would provide long-form data, it is relatively simple to determine the probability that any particular **household unit** would be chosen to be included in the long-form sample. In short, this probability of inclusion would be 1 in 5, or .20, for any household in group 1 described above and would be 1 for any household in groups 2 to 5 (since every household in these groups was to be chosen).

However, the units of analysis of interest to a researcher of the census data could be something other than households. The researcher might be interested, instead, in studying economic families, or census families[6] or persons. This means that the probabilities of inclusion of any of these other units also need to be calculated so that appropriate survey weights can be developed. For the census, these calculations are also straightforward because these other types of units of analysis are contained within households, and there was no sub-sampling done within households (i.e. all individuals in a household were to provide data if the household was chosen for the sample). Therefore the inclusion probability for any of these other units is just the same as that of their household. The inverses of these inclusion probabilities are called probability weights and in the census these are either 5 (for most of the population) or 1 (for the last four groups listed in Section 2 above).

As with most surveys, Statistics Canada does not recommend that this probability weight variable be the weight that is used for analysis with the census survey data in the long-form microdata files. In fact, Statistics Canada does not usually provide the probability weight variable on an analytical data file. Its use could result in biased estimates of quantities of interest. This is because problems are encountered in the collection of the census data—like in any other data-collection process. As an example, sometimes it is unclear whether a dwelling is occupied or not and thus whether a questionnaire should be left there, or it could be that a household receiving a long-form questionnaire does not provide responses to any questions past the basic questions that appear on a short-form questionnaire, or a household known to exist fails to provide any responses, or within a questionnaire the responses provided are inconsistent. Because of these and other problems, a complex process of editing and imputation takes place involving all the census data before the long-form census microdata files are produced for that census. The decision rules about what editing and imputation are done vary from census to

6. Since the definitions of census families and economic families have changed slightly over time, reference should be made to the documentation for the particular census being studied.

census, and determine the units to be included in the long-form microdata file. The probability weights of these units are then calibrated (or adjusted) to produce survey weights that, when used for estimation, will reduce or eliminate discrepancies between weighted sample estimates (using these survey weight variables) and population counts.

Since 1991, the probability weights have been adjusted to satisfy more than 30 constraints (obtained from the common questions on both the short and long forms) within each weighting area (WA).[7] Some constraints are at the household level while others are at the person level.[8] These calibrations adjust the probability weights for the types of people and households that are over- or under-represented amongst the records in the long-form data file. For example, one-person households and young males tend to be under-represented in the long-form file and would tend to have their probability weights adjusted upwards in the creation of survey weights. On the other hand, married people tend to be over-represented in the long-form file and thus would have their probability weights adjusted downwards. The methods used for weight calibrations have also changed somewhat from census to census as improved approaches have been developed; descriptions of the methods are in the technical reports available in the reference material available for each census.[9]

In the three most recent censuses (1996, 2001, and 2006) the weight adjustment process has been such that the survey weight variable to be used for person-level analysis is constant for all persons within the same household. This means that the **same weight variable can be used for analysis for any type of unit nested within households** – for analysis of households, for analysis of families or for analysis of individuals. As a consequence, only a single weight variable needs to be provided on the RDC microdata files for the 1996, 2001, and 2006 censuses.[10] For the 1996, 2001 and 2006 microdata files, as can be seen in Table 1, the name of this variable is **COMPW2**[11]. This is the survey weight variable that was used by Statistics Canada for published tabulations based on census sample data for those years.[12]

For the 1991 Census, the weight-adjustment process was such that the survey weight variable used by Statistics Canada for published tabulations at the person or family levels (called COMPW5 on the RDC data file) is not constant for all persons within the same household. Thus,

7. For the 2001 and 2006 Censuses, a WA consists, on average, of eight contiguous dissemination areas (DAs). The definition of WAs is a little different in the earlier censuses.

8. As an example of a person-level constraint, the weighted estimate of the number of married persons within each WA, using the long-form sample, was constrained to equal the population count of married persons in the WA obtained from all census questionnaires – both short and long.

9. For example, see the 1996 Census Technical Report, *Sampling and Weighting*. This report, and similar ones for the 2001 and 2006 censuses are available from the Statistics Canada website and are in the Reference section.

10. When earlier census files are released into the RDC's, some may contain more than one weight, with different weights being recommended for different units of analysis.

11. The RDC microdata file for the 2006 Census actually contains 2 weight variables – COMPW2 and COMPW1 – but both have the same value for records that have DocTp equal to values other than 4, 13 or 14. For records with these values of DocTp, COMPW1 has a "missing" value. (DocTp is a variable that gives a classification of households by type of census questionnaire that was used.)

12. Statistics Canada also produces what are called population estimates adjusted for net under-coverage, which are used by the federal government for transfer and equalization payments to provinces. These estimates are prepared after additional coverage studies are completed after the final survey weights have been calculated and thus a researcher cannot obtain these estimates through typical weighting procedures using the microdata files.

a different weight variable (called COMPW1 on the RDC data file) should be used for estimation at the household level than for estimation at person or family levels, if it is desirable to replicate published tabulations. Also, for estimation at the family level, the value of COMPW5 for a particular member of the family should be used, specifically, for estimation at the level of census family, the value of COMPW5 for the census family member for which the census family pointer variable CFPtr=0 should be the weight for the census family; similarly, for estimation at the level of economic family, the value of COMPW5 for the economic family member for which the economic family pointer variable EFPtr=0 should be the weight for the economic family.

Not all units have a survey weight that is different from their probability weight. All individuals in the groups 2 to 5 described in Section 2, where all households were chosen to receive a long form, have a survey weight of 1 — to show that they were chosen with certainty. Non-response in these groups was dealt with by approaches other than adjustments and calibrations to the probability weights. Documentation for each census could be consulted in order to determine how non-response in these groups was handled.

Table 1: Recommended Weight Variables for Different Units of Analysis

	Census Year			
Analysis Unit	**1991**	**1996**	**2001**	**2006**
Person	COMPW5	COMPW2	COMPW2	COMPW2
Census family	Value of COMPW5 of the census family member with CFPtr=0	COMPW2	COMPW2	COMPW2
Economic family	Value of COMPW5 of the economic family member with EFPtr=0	COMPW2	COMPW2	COMPW2
Household	COMPW1	COMPW2	COMPW2	COMPW2
Dwelling[1]	COMPW1	COMPW2	COMPW2	COMPW2 or COMPW1

Note: Recommended since they are the weight variables used for official Statistics Canada tabulations.
1. Also need restrictions to the dwellings for which dwelling variables of interest were collected.

What about dwelling-level analyses?

While the majority of questions on the long-form census were applicable to everyone, there were some questions for which this was not so, such as the questions about dwellings. As an example, dwelling questions are not applicable to people in overseas households or in collectives, among others. Thus, if an analysis at the level of the dwelling is to be done, the households for which the dwelling variables of interest make sense need to be identified, which is often possible through the use of the DocTp variable and knowledge of the contents of different types of questionnaires. The household weight variable can then be used as the suitable survey weight for the dwellings to be included in the analyses.

For 2006, two weight options are possible when doing dwelling-level analysis. On the one hand, the COMPW2 weight variable may be used, after restricting to households where the

dwelling variables of interest are collected. On the other hand, the COMPW1 weight variable is also an option since it has the same value as COMPW2 except for records with DocTp equal to 4 (Overseas Household 2C), 13 (Occupied senior units 2B) or 14 (Occupied Senior units 2D), for which COMPW1 has a "missing" value. Households with these 3 values of DocTp are among the households for which the dwelling variables were not applicable. Cautions have been issued, however, about the data quality of the DocTp variable. Analysts should refer to the 2006 census codebook (Statistics Canada, 2008) for more information.

For which censuses will one weight variable be sufficient?

As noted above, a single survey weight variable should suit the needs of analysts in the RDC's doing research with the census data files for 1996, 2001 and 2006. However, this will not be the case for analysis of earlier censuses. For the 1991 Census, as described above, and for censuses prior to 1991, different survey weights are needed for person-level and household-level estimates because the weight adjustments for persons were calculated independently from the weight adjustments for households.

Some researchers in the RDC's may have done analysis using the more extensive 1996, 2001 or 2006 census databases at Statistics Canada Head Office rather than using the files provided to the RDC's. They would have noted that these databases contain several weight variables, even though it is stated above that a single weight variable should be all that they require. If a researcher inspected these weight variables she would find that several would be the same – and equal the one on the RDC datafile, the others would have a constant value of 1. The inclusion of all these variables is a holdover from the time when these variables actually contained different values. It allows a researcher who is doing analysis with older censuses to repeat her estimation procedure for newer censuses, even down to the use of the same names for her survey weight variables.

Summary

Long-form census microdata files for 1991, 1996, 2001 and 2006 are accessible to researchers in the RDC's. These files, while containing person-level records, can also be used to study families, households and dwellings. Regardless of the unit of analysis, it is recommended that weighted techniques be used to produce the estimates of quantities of interest.

Because of the sampling plan for choosing those who were to fill out a long-form questionnaire and because of the approaches used to produce survey weights for analysis, the same weight variable can be used for producing estimates at the person, family, household or dwelling level for the 1996, 2001 and 2006 censuses. This means that the value of the weight variable for a particular individual in the person-level file is appropriate as a weight to represent the individual's family, household or dwelling, if an analysis was being done at one of those levels. On the other hand, for the 1991 Census, a different weight variable should be used for producing estimates at the person and family level than for producing estimates at the household and dwelling level. Table 1 indicates which weight variables to use so that Statistics Canada published tabulation values would be replicated for each of the four censuses.

Analysis does not usually stop with the production of weighted estimates of quantities of interest. Researchers usually wish to provide variability measures such as standard errors or confidence intervals or to carry out statistical tests. In order to do this appropriately, the clustering of individuals within households and sample stratification should be considered. A forthcoming paper will address these issues.

Acknowledgement

The author wishes to acknowledge the many contributions of Mike Bankier to the content of this article.

References

Statistics Canada. 1999. *1996 Census Technical Report: Sampling and Weighting*. Statistics Canada Catalogue no. 92-371-XIE. Ottawa, December 7. http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=92-371-X&CHROPG=1&lang=eng (accessed July 26, 2011).

Statistics Canada. 2004. *2001 Census Technical Report: Sampling and Weighting*. Statistics Canada Catalogue no. 92-395-XIE. Ottawa, December 15. http://www12.statcan.gc.ca/english/census01/Products/Reference/tech_rep/sampling/offline%20documents/92-395-XIE.pdf (accessed July 26, 2011).

Statistics Canada. 2008. *Research Data Centres (RDC): 2006 Census Code Book*. Census Operations Division, Statistics Canada, Ottawa, October 2008. http://www.statcan.gc.ca/dli-ild/meta/census-recensement/2006/rdc-codebook-public-2006-ver2-cdr-manuel-des-codes.pdf (accessed March 5, 2012).

Statistics Canada. 2009. *2006 Census Technical Report: Sampling and Weighting*. Statistics Canada Catalogue no. 92-568-X. Ottawa, August. http://www12.statcan.gc.ca/census-recensement/2006/ref/rp-guides/rp/sw-ep/sw-ep_index-eng.cfm (accessed July 26, 2011).

Instructions for authors

The Information and Technical Bulletin will accept submissions for articles that address methodological or technical topics related to the datasets that are available at the Research Data Centres.

Language of material

Manuscripts may be submitted in English or French. Accepted submissions will be translated into both official languages for publication.

Length of submissions

The maximum length of submitted articles should not exceed 20 pages, double-spaced, excluding programs and appendices. In addition to in-depth explanations of technical issues, the Bulletin also accepts short (3 page) submissions that provide quick solutions to analytical problems and commentary from fellow researchers about material previously released in the Bulletin.

File formats and layout of text

Manuscripts must be submitted in Microsoft Word (.doc) and may be sent by regular mail on a disk or CD or by email.

Manuscripts must have a cover page showing the names of the authors, their primary institution of affiliation, and the contact information (telephone number, mailing address and e-mail address) of the lead author.

Manuscripts must be prepared in 12pt Times New Roman, double-spaced, with 1-inch (2.5 cm) margins.

Titles should have sentence-case capitalization (e.g., Bootstrapping made easy…).

Boldface type should only be used for headings. Underlining and italics are not to be used for headings.

Footnotes and references should be single-spaced and formatted according to *The Canadian Style: A Guide to Writing and Editing*.

File formats and layout of tables and charts

Tables and charts must be submitted in Microsoft Excel worksheets (.xls) or in comma-separated value (.csv) format. Each file must be clearly named table1, chart6, etc.

Tables and charts may be sent by regular mail on a disk or CD, or by e-mail.

Do not insert tables or charts into the text, but indicate their location in the text by inserting the title, followed by the filename in parentheses, e.g.,

Chart 6. Chocolate consumption by children, Canada, 2000 (chart6)

Mathematical expressions

All mathematical expressions should be set out separate from paragraph text. Equations must be numbered, with the number appearing to the right of the equation flush with the margin.

Style guide

Please follow *The Canadian Style: A Guide to Writing and Editing*. It is available for purchase by contacting Government of Canada Publications, Public Works and Government Services Canada.

Address for submission

Manuscripts and all correspondence relating to the contents of the Bulletin should be sent to the

Editorial Committee

- by email to rdc-cdr@statcan.ca

The review process

The editorial committee conducts the initial article review process. Editors may solicit past authors of the Bulletin or subject matter experts to participate in the process. The articles submitted to the Bulletin are reviewed for accuracy, consistency, and quality.

Upon completion of the initial review, the articles undergo both peer and institutional review. Peer reviews are conducted in accordance with Statistics Canada's Policy on the Review of Information Products. Institutional reviews are be conducted by members of senior management within Statistics Canada in order to ensure that the material does not compromise the Agency's guidelines of standards, or reputation for non-partisanship, objectivity and neutrality.

For more information about the review process, please contact the Editorial Committee at the address above.